

# The $\beta$ -link motif in protein architecture

David P. Leader\* and E. James Milner-White

College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom.

\*Correspondence e-mail: david.leader@glasgow.ac.uk

Received 25 May 2021

Accepted 29 June 2021

Edited by B. Kobe, University of Queensland, Australia

**Keywords:**  $\beta$ -link;  $\beta$ -barrel;  $\beta$ -sandwich;  $\beta$ -bulge;  $\beta$ -turn; serine proteases; SARS-CoV-2 protease.

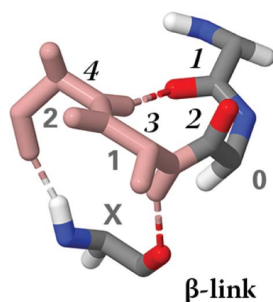
**Supporting information:** this article has supporting information at journals.iucr.org/d

The  $\beta$ -link is a composite protein motif consisting of a G1 $\beta$   $\beta$ -bulge and a type II  $\beta$ -turn, and is generally found at the end of two adjacent strands of antiparallel  $\beta$ -sheet. The 1,2-positions of the  $\beta$ -bulge are also the 3,4-positions of the  $\beta$ -turn, with the result that the N-terminal portion of the polypeptide chain is orientated at right angles to the  $\beta$ -sheet. Here, it is reported that the  $\beta$ -link is frequently found in certain protein folds of the SCOPe structural classification at specific locations where it connects a  $\beta$ -sheet to another area of a protein. It is found at locations where it connects one  $\beta$ -sheet to another in the  $\beta$ -sandwich and related structures, and in small (four-, five- or six-stranded)  $\beta$ -barrels, where it connects two  $\beta$ -strands through the polypeptide chain that crosses an open end of the barrel. It is not found in larger (eight-stranded or more)  $\beta$ -barrels that are straightforward  $\beta$ -meanders. In some cases it initiates a connection between a single  $\beta$ -sheet and an  $\alpha$ -helix. The  $\beta$ -link also provides a framework for catalysis in serine proteases, where the catalytic serine is part of a conserved  $\beta$ -link, and in cysteine proteases, including M<sup>pro</sup> of human SARS-CoV-2, in which two residues of the active site are located in a conserved  $\beta$ -link.

## 1. Introduction

The two major architectural features of proteins –  $\alpha$ -helix and  $\beta$ -sheet – occur where a suitable pair of dihedral angles is repeated along the backbone, leading to extensive hydrogen bonding. These features are often combined into higher order structures such as helix bundles,  $\beta$ -sandwiches and  $\beta$ -barrels. At the opposite level of complexity are small (three- to six-residue) structural elements – protein motifs – that are defined by combinations of dihedral angles and patterns of hydrogen bonding. Such motifs may cause changes in the direction of the polypeptide chain [for example, in the  $\beta$ -turn (Venkatachalam, 1968; Richardson, 1981; Wilmot & Thornton, 1988; Hutchinson & Thornton, 1994; Gunasekaran *et al.*, 1998) and the Schellman loop (Schellmann, 1980; Milner-White, 1988)] or produce indentations in protein surfaces that enable them to bind particular ligands [for example, in the nest (Watson & Milner-White, 2002; Afzal *et al.*, 2014) and the crown bridge (Leader & Milner-White, 2015)].

The  $\beta$ -bulge is a small motif in which a single residue (X) on one  $\beta$ -strand forms hydrogen bonds to two successive residues (1 and 2; referred to as the ‘doubleton’) of a second, usually antiparallel,  $\beta$ -strand. Two broad categories of  $\beta$ -bulge have been defined: classic  $\beta$ -bulges, which have the  $\alpha_R$  conformation at position 1, and G1  $\beta$ -bulges, with the  $\alpha_L$  conformation at this position (Richardson *et al.*, 1978). We have recently shown that two classes of G1  $\beta$ -bulge can be distinguished on the basis of the conformation of the residue (numbered 0) preceding the doubleton: G1 $\alpha$ , where the conformation is  $\alpha_R$ , and G1 $\beta$ , where it is  $\beta_R$  (Leader & Milner-White, 2021). This led to the recognition that the members of each class of G1



OPEN ACCESS

$\beta$ -bulge most frequently occur as part of a composite with a specific second small hydrogen-bonded motif. In the case of the G1 $\alpha$   $\beta$ -bulge this composite is with a type I  $\beta$ -turn and is the well known  $\beta$ -bulge loop (Richardson *et al.*, 1978; Milner-White, 1987; Chan *et al.*, 1993; Blandl *et al.*, 2003; Craveur *et al.*, 2013). Almost all G1 $\alpha$   $\beta$ -bulges (99%) are found in  $\beta$ -bulge loops (Leader & Milner-White, 2021).

In the case of the G1 $\beta$   $\beta$ -bulge the composite is with a (hydrogen-bonded) type II  $\beta$ -turn, with the 1,2-positions of the  $\beta$ -bulge corresponding to the 3,4-positions of the  $\beta$ -turn (Fig. 1). The role of the  $\beta$ -turn in the composite can be regarded as making a hydrogen bond to ensure that the polypeptide chain not only enters or leaves the  $\beta$ -sheet at the position of the disruptive  $\beta$ -bulge, but that it does so in a direction that is perpendicular to the  $\beta$ -sheet. In this respect, it differs from most other small hydrogen-bonded motifs, which result in either turns or indentations in proteins. This composite motif has not received much attention since it was originally described (Richardson *et al.*, 1978). It has recently been assigned a name –  $\beta$ -link – and is the subject of the present work.

The  $\beta$ -link includes 85% of G1 $\beta$   $\beta$ -bulges (Leader & Milner-White, 2021), and approximately 22% of proteins contain at least one instance of this motif. As there are fewer  $\beta$ -links in proteins than  $\beta$ -sheets, the question arose as to whether the motif only occurred in particular structural

contexts. This has been approached here by first identifying which folds of the SCOPe classification of protein architecture contain  $\beta$ -links and then determining whether the motifs occupy specific positions within these folds. We report that the majority of  $\beta$ -links are found where an antiparallel  $\beta$ -sheet connects to certain other structural domains. Two situations predominate: in certain  $\beta$ -sandwich structures at the junction between the two component sheets, and in many small  $\beta$ -barrels at one end of a loop across an end of the barrel. Thus, the role of the  $\beta$ -link is not merely to provide a terminus to a  $\beta$ -sheet, but to form a connection between specific components of larger units of protein architecture.

## 2. Materials and methods

This work employed a new MySQL relational database – Protein Motif 2 – containing structural information from a set of 4484 individual protein subunits derived from the Richardson laboratory Top 8000, which is 70% nonredundant in terms of structure (<http://kinemage.biochem.duke.edu/databases/top8000.php>). It was populated with small hydrogen-bonded motifs, using the program *HBPlus* (McDonald & Thornton, 1994) to determine the hydrogen bonds, in a similar manner to that described previously for an original, smaller, database (Leader & Milner-White, 2009). Protein Motif 2 incorporates a table listing the SCOPe clas-

sification (Murzin *et al.*, 1995; Fox *et al.*, 2014; Chandonia *et al.*, 2017) of the fold of the protein in which individual motifs occur. The data for this were derived from the file `dir.cla.scop.2.07-stable.txt` from the SCOPe web resource (<http://scop.berkeley.edu>). A web application (*Motivated Proteins 2*) utilizing this database to allow queries on the basis of SCOPe fold is now publicly available at <http://motif.gla.ac.uk/ProtMotif21/index.html>. [Version 2 of the free desktop tool *Structure Motivator* (Leader & Milner-White, 2012) also incorporates this new database.]

The  $\beta$ -links in the database were determined from SQL queries of the Protein Motif 2 database for structures with the three hydrogen bonds shown in the inset in Fig. 1, together with a requirement that the dihedral angles of residue 2 be within the  $\beta_R$  region of the Ramachandran plot. This selects for the combination of a type G1 $\beta$   $\beta$ -bulge and a type II  $\beta$ -turn shown in CPK colouring in Fig. 1(d) without any separate explicit queries for these constituents. It should be emphasized that the specification does not include the two hydrogen bonds between the two  $\beta$ -strands shown in white in Fig. 1(d), so that  $\beta$ -links are

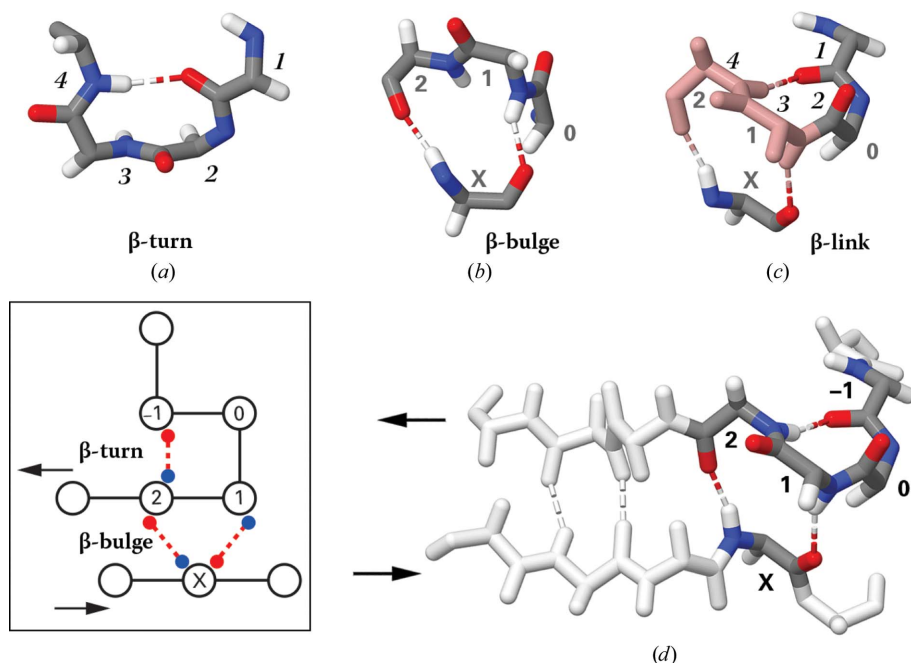


Figure 1

Structure of the  $\beta$ -link and its relationship to the constituent motifs. (a) Type II  $\beta$ -turn. (b) G1 $\beta$   $\beta$ -bulge. (c)  $\beta$ -Link. The residues are numbered either as for the  $\beta$ -turn (italic serif font) or the  $\beta$ -bulge (roman sans-serif font) or both. The two hydrogen-bonding residues of the  $\beta$ -bulge (3 and 4) that are common to the  $\beta$ -turn (1 and 2) are coloured pink. (d) The  $\beta$ -link in the context of two antiparallel strands of a  $\beta$ -sheet in staphylokinase (PDB entry 2sak; Rabijns *et al.*, 1997). The residues of the  $\beta$ -link are represented using the CPK colour scheme with hydrogen bonds coloured red. Proximal residues and hydrogen bonds are coloured white. The numbering is as for the  $\beta$ -bulge, extended in the negative direction. Inset: diagrammatic representation of the hydrogen-bonding pattern with residues numbered as in (d).

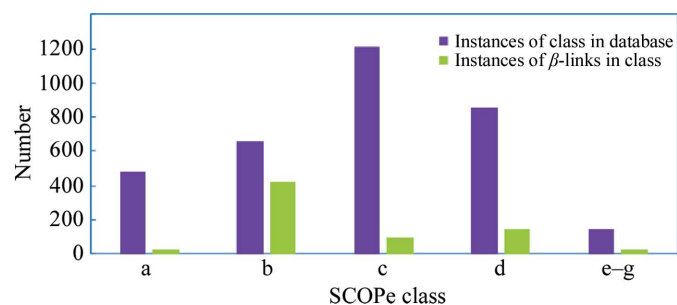
not necessarily located at the ends of perfect antiparallel  $\beta$ -strands.

The numerical data reported here (Fig. 2 and supporting information) are from the analysis of  $\beta$ -links in the Protein Motif 2 database. Of the 4484 protein subunits in the database, only 2885 had an assigned SCOPe ID in release 2.07. Of the total of 1283  $\beta$ -links, 973 resided in a protein fold that had been assigned a SCOPe ID, and these constituted the subset analysed. Further work to determine the extent of conservation of  $\beta$ -links in particular folds was performed by examining other proteins with the same or related IDs using the SCOPe web resource. Positions in these proteins corresponding to  $\beta$ -links in the database were identified either by visual inspection using the 3D protein graphics program *Jmol* (Herráez, 2006) or, where necessary, with the multiple sequence alignment program *ClustalX* (Larkin *et al.*, 2007).

### 3. Results

#### 3.1. Assignment of $\beta$ -links to architectural components of proteins

SCOPe (Structural Classification Of Proteins; Fox *et al.*, 2014) was used to determine whether  $\beta$ -links occur in particular types of protein domain. This classification has the hierarchy 'class' (designated by a character) followed by 'fold',



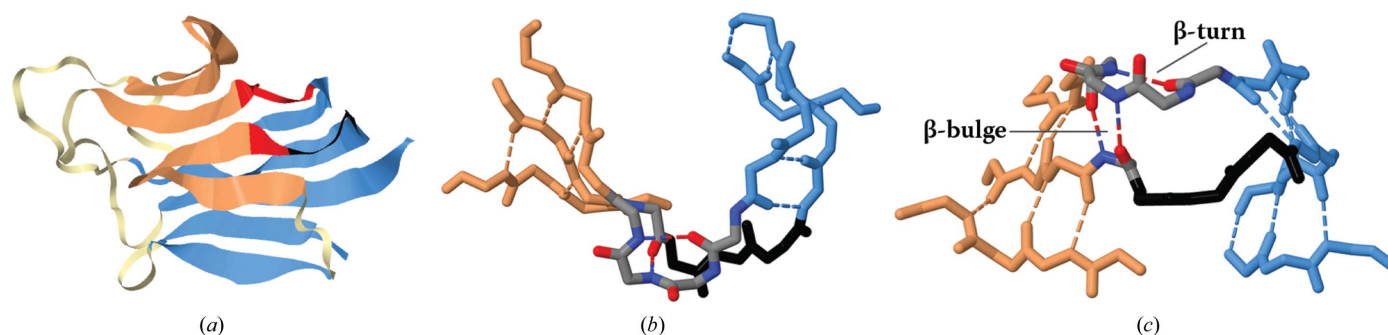
**Figure 2**  
Distribution of  $\beta$ -links between different SCOPe classes. Purple bars: number of instances of the class in the database. Green bars: number of  $\beta$ -links in the class. Note that in some cases there is more than one  $\beta$ -link in a particular instance of a SCOPe class.

'superfamily' and 'family' (designated by numerals). For example, in group b.1.1.1 the class is 'All beta proteins', the fold is 'Immunoglobulin-like beta-sandwich', the superfamily is 'Immunoglobulin' and the family is 'V-set domains'. The SCOPe ID for the environment of each  $\beta$ -link was identified and the great majority were found to be in classes b, c or d. The distribution of  $\beta$ -links is seen in Fig. 2. We initially examined folds of class b, which contain 58% of all  $\beta$ -links. The class comprises 178 folds (b.1, b.2 *etc.*), although many have few members. This was simplified by considering the 12 folds described as  $\beta$ -sandwich or  $\beta$ -sandwich-like as one group, including 18% of all  $\beta$ -links, and the 15 folds described as  $\beta$ -barrels as another group, containing 22% of all  $\beta$ -links. Distributions of  $\beta$ -links between folds are available in Supplementary Fig. S1 and Supplementary Table S1.

#### 3.2. $\beta$ -Links in sandwich and sandwich-like structures

$\beta$ -Sandwich proteins (folds b.1–b.33) consist of two antiparallel  $\beta$ -sheets, the planes of which are juxtaposed and in many cases linked by one or more strands of polypeptide chain. The double-stranded  $\beta$ -helix (fold b.82) is similar in these respects, and we regard it as  $\beta$ -sandwich-like. Together, these folds constitute 46% of all class b folds, in the database which include an approximately similar proportion of the  $\beta$ -links in class b. It emerged that many of these  $\beta$ -links have the role, which has not previously been reported, of making an angular connection between the two sheets of a  $\beta$ -sandwich.

Fig. 3 illustrates three different aspects of the connection. The sharp bend between the two sheets is evident in the ribbon view of the whole domain, in which the  $\beta$ -link is towards the upper right of one sheet (salmon) and makes a connection to a second sheet (blue). The  $\beta$ -link makes this connection through the N-terminal portion of the strand on which the doublet of the  $\beta$ -bulge is located (Fig. 3a). A plan view of three strands in the proximity of the  $\beta$ -link shows how the planes of the  $\beta$ -turn and the left-hand (salmon) sheet are approximately at right angles (Fig. 3b). More precisely, the first (N-terminal) residue of the  $\beta$ -turn, that nearest to the right-hand (blue) sheet, is directed away from the right-hand sheet in an approximately perpendicular orientation to it.



**Figure 3**  
Geometry of the junction in a  $\beta$ -sandwich at the position of a  $\beta$ -link. Illustrated is jelly-roll/ $\beta$ -sandwich SCOPe domain d2q1ma1 of glucocorticoid-induced TNF receptor ligand (PDB entry 2q1m; Chattopadhyay *et al.*, 2007), SCOPe family b.22.1.1. (a) Ribbon plot of the domain. The  $\beta$ -strands of the two sheets are coloured salmon and light blue, the  $\beta$ -link is red and the connection made by the strand with the singleton is black. Other non- $\beta$ -strand regions are coloured cream. (b) The backbone atoms of the two strands of the  $\beta$ -link are shown, together with the adjacent strand, in a plan view. (c) As (b) but in a side view, with constituent motifs of the  $\beta$ -link indicated to allow comparison with Fig. 1.

Here, the hydrogen bonding of the two upper strands of the right-hand sheet extends to the start of the  $\beta$ -turn, which is the tightest situation possible (Fig. 3c). However, such close relative positioning is not found in every  $\beta$ -sandwich.

The number and topology of the strands vary considerably between different  $\beta$ -sandwich and  $\beta$ -sandwich-like protein folds, as does the position of the  $\beta$ -link. Their common feature is that the  $\beta$ -link is at the effective corner of the sheet on which it resides, even when it is not actually on the first strand. Fig. 4 uses diagrams (Figs. 4a–4f) to show how this idea can accommodate  $\beta$ -sandwich proteins which show increasing divergence from a situation of two ‘ideal’ sheets with the same number of antiparallel strands of similar length. Proteins illustrating the successive features introduced in Figs. 4(a)–4(f) are presented in Figs. 4(g)–4(l), although it is emphasized that in other respects (for example the total number of

strands) these proteins may not correspond to the simple patterns in the diagrams.

We start with two similar sheets in which antiparallel strands are linked by the continuation of the two strands on which the  $\beta$ -link (red) occurs (Fig. 4a). This is illustrated by  $\alpha$ -amylase inhibitor (Fig. 4g), which has three strands in each sheet rather than the four depicted in Fig. 4(a). One variation of this involves a different number of strands in the two sheets (Fig. 4b), where the antiparallel nature of the sheet at the back is disrupted by an additional strand (dark grey). This is illustrated by CD58 (Fig. 4h). In another variation (Fig. 4c) the  $\beta$ -link is not the first (‘upper’) strand. Here, an additional strand (dark grey) is shown above the linking strand, displaced from the position of the  $\beta$ -link. It is illustrated for the TNF receptor ligand (Fig. 4i). In some cases the second strand from the front sheet is not connected to the sheet behind (Fig. 4d),

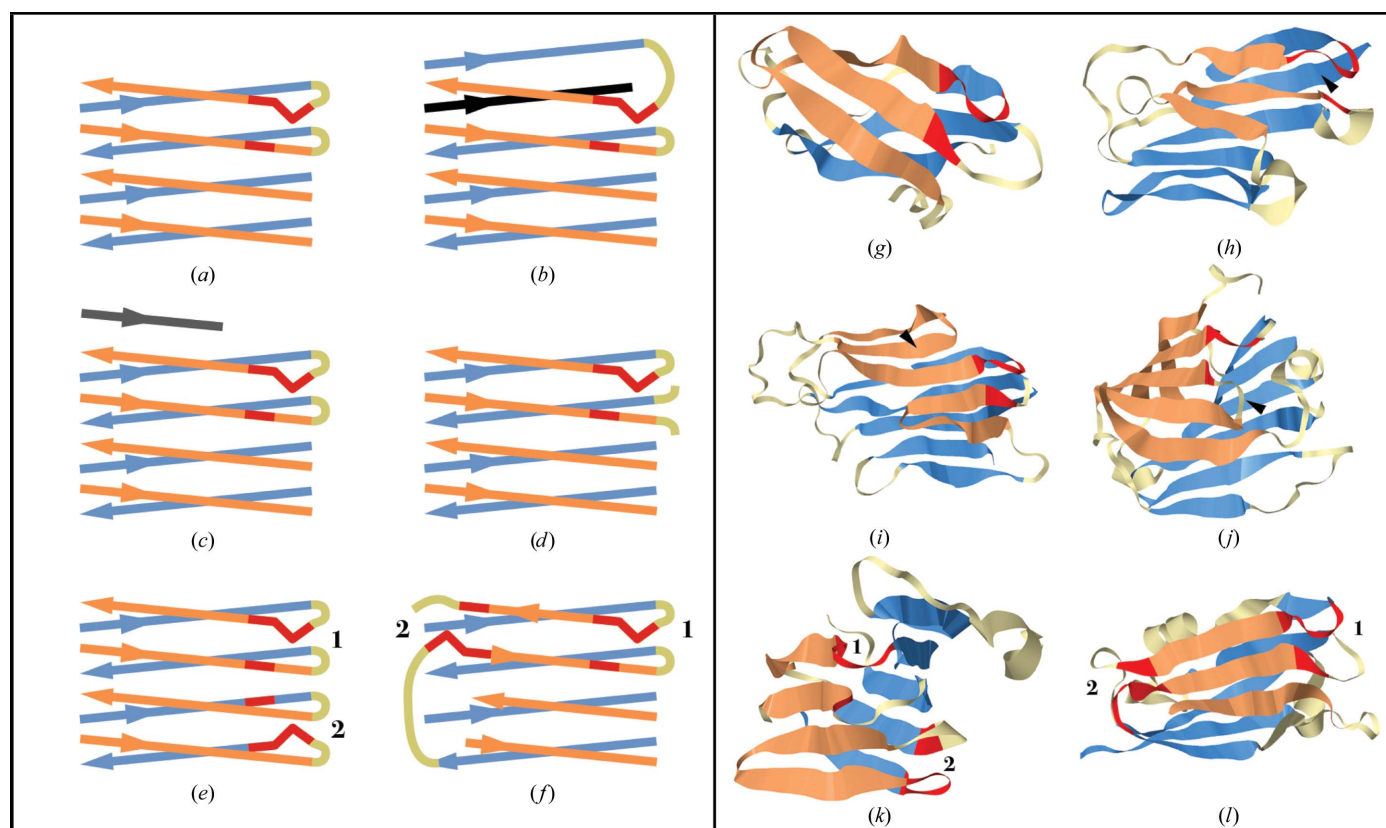


Figure 4

Arrangements of strands in some  $\beta$ -sandwiches containing  $\beta$ -links. Diagrammatical representations are shown (a–f), with corresponding illustrations as ribbon models (g–l). In the diagrams the two layers of the sandwich are coloured salmon and blue, while the  $\beta$ -links are red: the  $\beta$ -turn is indicated by a kink and the line representing the singleton of the  $\beta$ -bulge is shorter. In the ribbon models the colour scheme is similar to that in Fig. 3(a). (a) The simplest situation, in which each sheet has the same number of antiparallel strands and the two upper strands are connected between both sheets. (b) As (a), but with an additional strand (dark grey) in the back sheet separating the two upper strands that are adjacent in the front sheet. (c) As (a), but with an additional shorter strand (dark grey) above what was previously the upper strand. (d) Second (singleton) strand of the motif not connected directly to the back sheet. (e) With a second  $\beta$ -link (2) oriented in the opposite direction on the back sheet. (f) With a second  $\beta$ -link (2) lying between the same pair of strands in the same sheet as the first  $\beta$ -link, but oriented in the opposite direction. (g) Tendamistat, a bacterial  $\alpha$ -amylase inhibitor (PDB entry 1bvn; Wiegand *et al.*, 1995); SCOPe domain d1bvt\_, b.5.1.1. (h) CD2 binding domain of CD58 (PDB entry 1ccz; Ikemizu *et al.*, 1999); SCOPe domain d1ccza1, b.1.1.1. The additional strand corresponding to that in (b) is marked with an arrowhead. (i) Charcot–Lyden crystal protein (PDB entry 1lcl; Leonidas *et al.*, 1995); SCOPe domain d1lcl\_, b.29.1.3. The additional strand corresponding to that in (c) is marked with an arrowhead. (j) TNF receptor ligand (PDB entry 2q1m; Chattopadhyay *et al.*, 2007); SCOPe domain d2q1ma1, b.22.1.1. The continuation of the second strand of the  $\beta$ -link is marked with an arrowhead. (k) Probable antibiotic synthesis protein TTHA0104 (PDB entry 1v70; Yokoyama *et al.*, 2008); SCOPe domain d1v70a\_, b.82.1.9. (l) Plastocyanin (PDB entry 3cvb; Crowley *et al.*, 2008); SCOPe domain d3cvba\_, b.6.1.1.



as illustrated by Charcot–Lyden crystal protein (Fig. 4*j*), where it continues to the strand below.

Certain  $\beta$ -sandwich or  $\beta$ -sandwich-like folds contain two  $\beta$ -links. One such situation is that shown in Fig. 4(*e*), where the second  $\beta$ -link ('2') is on the back sheet, with pseudo-symmetry to that on the front sheet. We observed this in the double-stranded  $\beta$ -helix, an example of which is shown in Fig. 4(*k*).

The other situation is a feature of cupredoxin-like folds of type b.6.1 and has a second  $\beta$ -link diametrically opposite to the first at the other end of the same pair of  $\beta$ -strands (Fig. 4*f*). This may seem surprising in relation to disruption of hydrogen bonding between strands 2 and 3 but, as the diagram indicates and the example of plastocyanin illustrates (Fig. 4*l*), strand 3 is always truncated (*cf.* Fig. 4*c*). (Fold b.6.1 only has three strands on the front sheet.) It is possible to consider this  $\beta$ -link as being at the bottom left of the front sheet.

In all but one of the folds where the  $\beta$ -link does occur it is found in the context of the two flanking hydrogen bonds shown in white in Fig. 1(*d*). An exception is fold b.1.1 (immunoglobulin-like), where the apparent insertion of an extra residue results in the flanking hydrogen bonds being separated from the  $\beta$ -link by a wide  $\beta$ -bulge. This is shown in Figs. 5(*a*) and 5(*c*), with the corresponding region of a b.1.2 fold included for comparison (Figs. 5*b* and 5*d*).

$\beta$ -Links are absent from two folds (b.3 and b.11) classified as  $\beta$ -sandwich-like for which there is significant representation in the database (over four members; see Supplementary Table S1). In all five examples of fold b.11, each with two similar

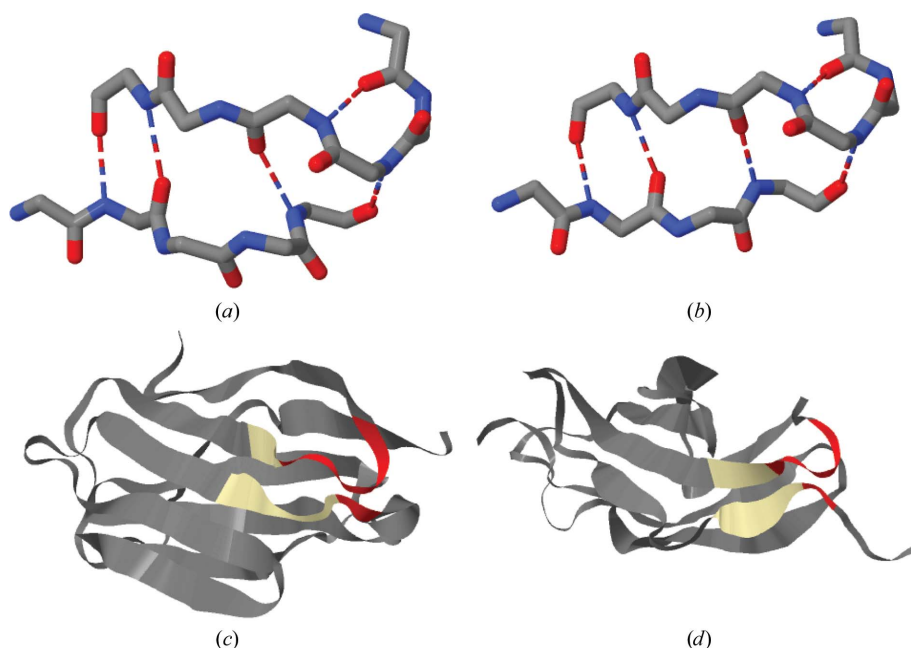
domains, connections between the two sheets are associated with (and are presumably facilitated by) a combination of a classic  $\beta$ -bulge and an adjacent glycine residue (Supplementary Fig. S2). In contrast, the connections in the nine examples of fold b.3 are not associated with any single structure, but exhibit a variety of  $\beta$ -bulges of different types or none at all.

Of 263  $\beta$ -sandwich-like proteins in the database (folds b.1–b.33 and fold b.82), 49% contain at least one  $\beta$ -link, with there being 180  $\beta$ -links in total. We have inspected each of these  $\beta$ -links and have ascertained that 99% lie at a corner between the two sheets of the sandwich and 91% are involved in a direct connection between the corners of the two sheets of the types shown in Fig. 4. The group of 8% that are not involved in a direct connection contains ten with a very extended connection and five in which the connection is between two adjacent strands on the same sheet.

### 3.3. $\beta$ -Links in small $\beta$ -barrels

Approximately 40% of the  $\beta$ -links found in SCOPe class b folds in the database occur in  $\beta$ -barrels (folds b.34–b.62). Of these, over 90% have 4–6 strands and are categorized as small (Youkharibache *et al.*, 2019), with most of the remainder being eight-stranded barrels with atypical strand arrangements (Supplementary Fig. S1). The striking, and previously unreported, feature of most of these  $\beta$ -links is that they make a connection from the N-terminus of one strand of the  $\beta$ -barrel, through a polypeptide loop, to the C-terminus of another strand of the same  $\beta$ -barrel. This and other aspects of the connection are illustrated in Fig. 6 for a variety of different types of  $\beta$ -barrel.

An example,  $\alpha$ -spectrin, of the large group of four-stranded  $\beta$ -barrels containing a  $\beta$ -link is shown in Fig. 6(*a*). The antiparallel nature of the strands in four-stranded barrels means that a connection across either of the two ends of the barrel is only made between adjacent strands (Fig. 6*a*, i). In contrast to the  $\beta$ -bulge loop (which is also found connecting strands in  $\beta$ -barrels) this is not a tight (5–6-residue) connection: in this example it is a loop of 14 residues crossing the upper end of the barrel (Fig. 6*a*, ii).  $\beta$ -Links in barrels of this fold (b.34) are flanked by the two additional hydrogen bonds in Fig. 1(*d*), as are  $\beta$ -links of most other  $\beta$ -barrel folds (Supplementary Table S1). The exception is fold b.36 (PDZ domain-like), where they are generally absent or, at best, weak. A typical example is shown in Fig. 6(*b*), where it can be seen that the divergence of the two strands only allows one hydrogen bond (marked with an arrowhead).



**Figure 5**

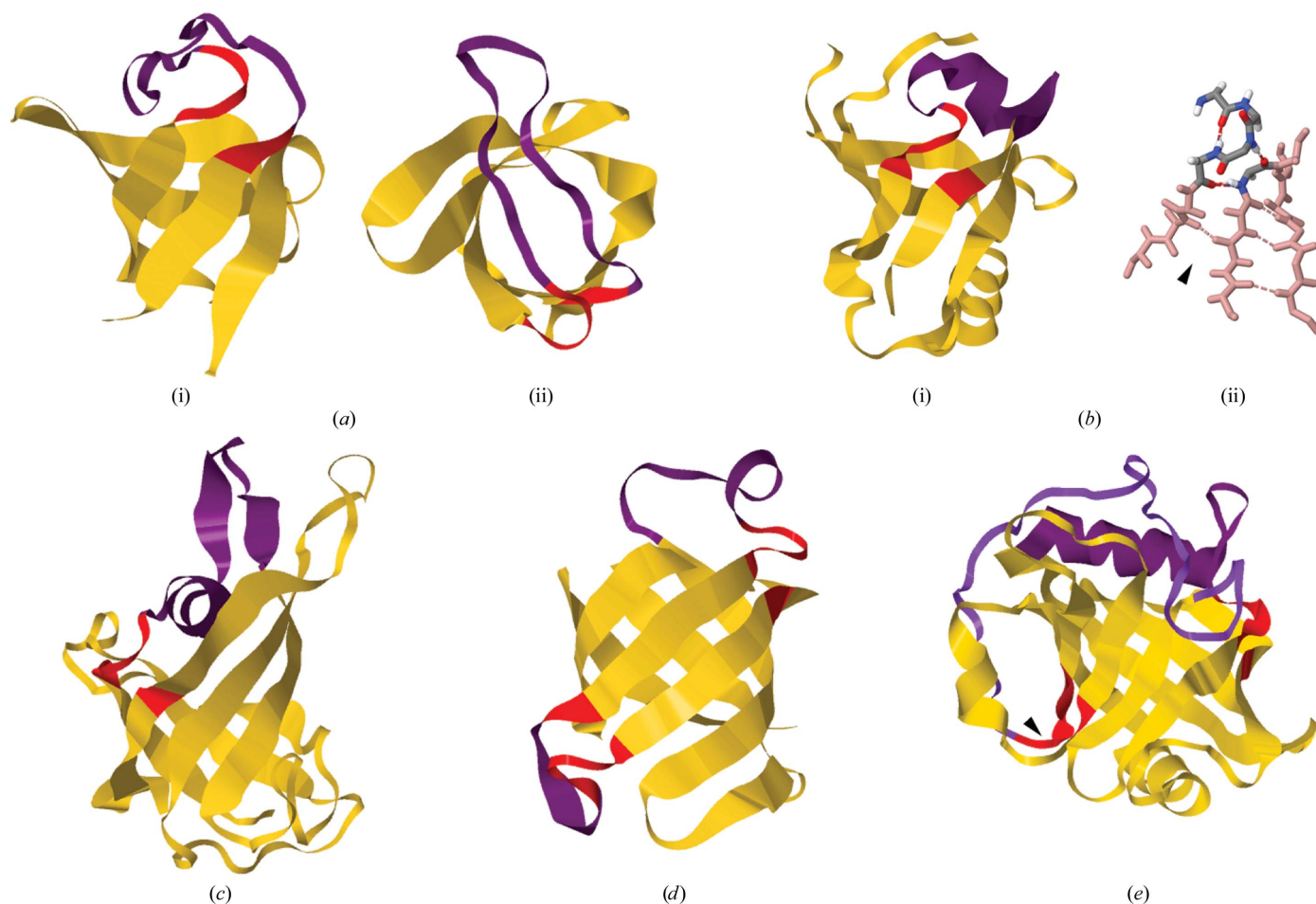
Unique features of the  $\beta$ -link in the immunoglobulin fold. (*a*) Region of the main chain in the vicinity of the  $\beta$ -link in T-cell receptor V $\alpha$ 11 (PDB entry 1h5b; Machius *et al.*, 2001); SCOPe domain d1h5bc\_, family b.1.1.1.1. (*b*) Region of the main chain in the vicinity of the  $\beta$ -link in the FnIII tandem A77–A78 domain of titin (PDB entry 3lpw; Bucher *et al.*, 2010); SCOPe domain d3lpwb1, family b.1.2.0. (*c*) Fold as in (*a*), but shown complete in ribbon representation. The  $\beta$ -link is coloured red, and the rest of the region shown in (*a*) is coloured cream. (*d*) Fold as in (*b*), but shown complete for a single repeat in ribbon representation. The colour scheme is as in (*c*).

An example of a  $\beta$ -link in a five-stranded  $\beta$ -barrel is seen in Figs. 7(c) and 7(d) (OB-fold protein, eIF5a). It connects a short  $3_{10}$ -helix across the end of the barrel. A  $\beta$ -link in a six-stranded antiparallel  $\beta$ -barrel is shown in Fig. 6(c) for ferredoxin reductase. Like many such six-stranded  $\beta$ -barrels, the strand connecting to the  $\beta$ -turn of the  $\beta$ -link is directly opposite, three strands away. In this example the connecting polypeptide is extensive, including a short helix across the 'upper' end of the barrel. A different six-stranded  $\beta$ -barrel, in which not all strands are antiparallel, is the double- $\psi$   $\beta$ -barrel. The example in Fig. 6(d), endoglucanase V, contains two  $\beta$ -links: one at each end of the barrel.

Fig. 6(e) shows the two  $\beta$ -links in the eight-stranded  $\beta$ -barrel of cyclophilin (b.62), which is unusual in that some strands are parallel. The  $\beta$ -link at the top of the image resembles those in the smaller  $\beta$ -barrels by making a connection to the opposite side across the 'upper' end of the

barrel by a three-turn  $\alpha$ -helix. The topology of the b.62 fold differs from other eight-stranded  $\beta$ -barrel folds (b.60 and b.61), in which the regular meander is expected to preclude this type of  $\beta$ -link. The second  $\beta$ -link in Fig. 6(e) is atypical in being directed away from the barrel (see below).

The location of a  $\beta$ -link within the  $\beta$ -sheet of a  $\beta$ -barrel differs from that in a  $\beta$ -sandwich, where occurrence at the end of an outer strand involves no loss of hydrogen bonding. The situation in a  $\beta$ -barrel with an even number of strands is that it is an antiparallel  $\beta$ -sheet rolled into a cylinder in which every strand is hydrogen-bonded to two other strands. However, the strands of the  $\beta$ -barrel are staggered with respect to one another, so that the extremities of a strand are only hydrogen-bonded to one adjacent strand. Thus,  $\beta$ -links can occur in one direction without disrupting the hydrogen bonding of the  $\beta$ -barrel, but not in the other direction. This is illustrated in Fig. 7(a), where two  $\beta$ -links (red) are shown at the top: one in



**Figure 6**

Ribbon models of domains of  $\beta$ -barrels containing  $\beta$ -links. Barrels are coloured gold,  $\beta$ -links are coloured red and the polypeptide chain between the  $\beta$ -turn of the  $\beta$ -link and the connecting  $\beta$ -strand is coloured purple. (a) Four-stranded, partly open,  $\beta$ -barrel:  $\alpha$ -spectrin (PDB entry 1bk2; Martinez *et al.*, 1998); SCOPe domain d1bk2a\_, family b.34.2.1. (i) View from the side of the barrel. (ii) View from the 'top' of the barrel. (b) Four-stranded, partly open,  $\beta$ -barrel: PDZ domain of human pick1 (PDB entry 2gzv; Elkins *et al.*, 2007); SCOPe domain d2gzva1, family b.36.1.0. (i) Ribbon representation. (ii) Backbone detail with the  $\beta$ -link coloured CPK and adjacent parts of the strands coloured salmon pink. The arrowhead indicates the position of the single flanking hydrogen bond. (c) Six-stranded, closed,  $\beta$ -barrel: ferredoxin reductase (PDB entry 1fnc; Bruns & Karplus, 1995); SCOPe domain d1fncal, family b.43.4.2. (d) Double- $\psi$ , closed, six-stranded  $\beta$ -barrel: endoglucanase V (PDB entry 2eng; Davies *et al.*, 1995); SCOPe domain d2eng\_, family b.52.1.1, with two  $\beta$ -links. (e) Eight-stranded, closed  $\beta$ -barrel: cyclophilin (PDB entry 2cpl; Ke, 1992); SCOPe domain d2cpla\_, SCOPe family b.62.1.1. It has two  $\beta$ -links; the nonstandard one is marked with an arrowhead.

one orientation and one in the other. In each case the strand adjacent to that with the doubleton is shown in green, and the potential hydrogen bonding that the residues N-terminal to the doubleton would make to that strand is indicated by the grey gradient fill. At position '1' no hydrogen bonds are made in any case, so the  $\beta$ -turn in the  $\beta$ -link does not disrupt hydrogen bonding. At position '2' the hydrogen-bonding is disrupted. This is why a  $\beta$ -link at position '2' is rarely observed. An exception was mentioned above for cyclophilins (Fig. 6e) and is shown in more detail in Supplementary Fig. S3.

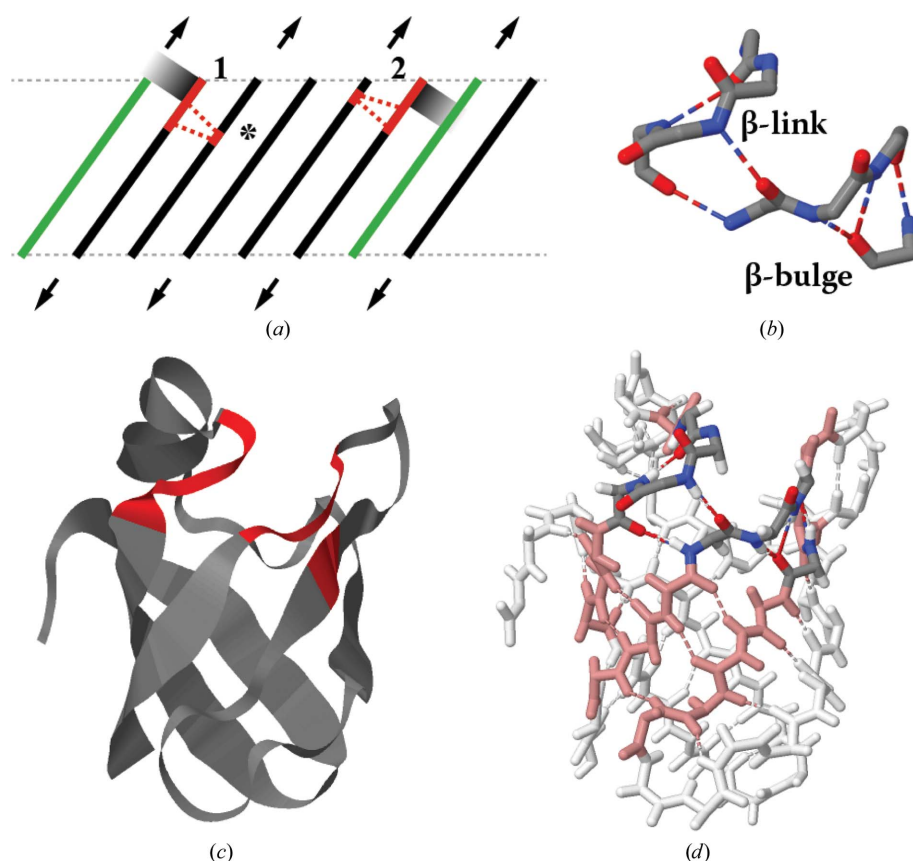
The  $\beta$ -links in several classes of  $\beta$ -barrel exhibit an additional structural feature which involves hydrogen bonding to a third strand. If one considers the second strand to be that on which the singleton of the G1 $\beta$   $\beta$ -bulge of the  $\beta$ -link is situated, an additional  $\beta$ -bulge of the 'classic' type, with three hydrogen bonds, is made between this and a third, antiparallel, strand. This is shown in Fig. 7(b), where it is seen that positions 1 and 2 of the classic  $\beta$ -bulge are  $X + 1$  and  $X + 2$  in relation to

the G1 $\beta$   $\beta$ -bulge. Figs. 7(c) and 7(d) show this in the context of the whole of the  $\beta$ -barrel. This additional classic  $\beta$ -bulge associated with the  $\beta$ -links is a feature of many small barrels, particularly in the PDZ domain (Fig. 6b), which functions to bind specific proteins. This involves the third strand, above, extending the  $\beta$ -sheet to a fourth strand: the C-terminal peptide of the ligand (see Supplementary Fig. S4).

The additional classic  $\beta$ -bulge is also found in the double-stranded  $\beta$ -helix, but is rare in the  $\beta$ -sandwiches.  $\beta$ -Links are absent from three folds (b.50, b.55 and b.61) classified as  $\beta$ -barrels for which there is significant representation in the database, and there are only a few  $\beta$ -links in folds b.45 and b.60: see Supplementary Table S1.

Of 219 proteins in the database with small  $\beta$ -barrels (folds b.34–b.62), 60% contain at least one  $\beta$ -link, with there being 167  $\beta$ -links in total. We have inspected each of these  $\beta$ -links and have ascertained that 92% are formed between adjacent strands at the end of the barrel and 76% form connections

across the end of the barrel of the types shown in Fig. 6. The group of 16% that are appropriately located but are not involved in a connection across the end of the barrel include some that connect to other parts of the protein and some where the connecting chain terminates above the barrel.



**Figure 7**

Structural aspects of  $\beta$ -links in  $\beta$ -barrels. (a) Diagram of the  $\beta$ -sheet of a  $\beta$ -barrel opened out and flattened, showing the stagger of the strands. The  $\beta$ -links are shown in red, with the hydrogen bonds of the  $\beta$ -bulge component represented as red broken lines. The second strand adjacent to the strand containing the  $\beta$ -turn of the  $\beta$ -link is shown in green, and the presence (2) or absence (1) of potential hydrogen bonding to it is indicated by the grey gradient projected from the N-terminal two residues of the  $\beta$ -turn. The two  $\beta$ -links have been placed as if both were at the top of the barrel. The asterisk indicates the position of the frequently associated classic  $\beta$ -bulge, shown in (b). (b) Hydrogen-bonding pattern of the combination of  $\beta$ -link and classic  $\beta$ -bulge found in many  $\beta$ -barrels. (c) Ribbon diagram showing the position of the  $\beta$ -link/ $\beta$ -bulge structure, coloured red, in the five-stranded barrel of eIF5a [PDB entry 1bkb; Peat *et al.*, 1998; also for (b)]: SCOPe domain d1dkba2, SCOPe family b.40.4.5. (d) As in (c), but a backbone diagram with the  $\beta$ -link/ $\beta$ -bulge combination coloured CPK, the remaining parts of the three strands on which they occur salmon pink, and the rest of the  $\beta$ -barrel white.

### 3.4. Other $\beta$ -links

SCOPe class c and d folds contain mixtures of  $\alpha$ -helices and  $\beta$ -sheets. Although the proportion of  $\beta$ -links in c and d folds is low, together they comprise 34% of  $\beta$ -links in the database (Fig. 2). Many fall into specific folds: in three or four (c.66, d.26, d.145 and perhaps d.157) the  $\beta$ -link occurs where a strand of a  $\beta$ -sheet connects to an  $\alpha$ -helix. Two examples are shown in Supplementary Fig. S5. In both, the  $\beta$ -link is associated with a classic  $\beta$ -bulge as in Fig. 7(b).

Not all  $\beta$ -links are found in the structural environments discussed above, and instances occur in various situations, some of which do not involve  $\beta$ -strands. Two structures, seen in Supplementary Fig. S6, involving a pair of cooperating  $\beta$ -links, are worthy of attention. The first is fold b.85, the  $\beta$ -clip (Iyer & Aravind, 2004), which is what might be regarded as a belt of two  $\beta$ -strands looped out from the rest of the structure. The second is fold b.84, the barrel-sandwich hybrid.



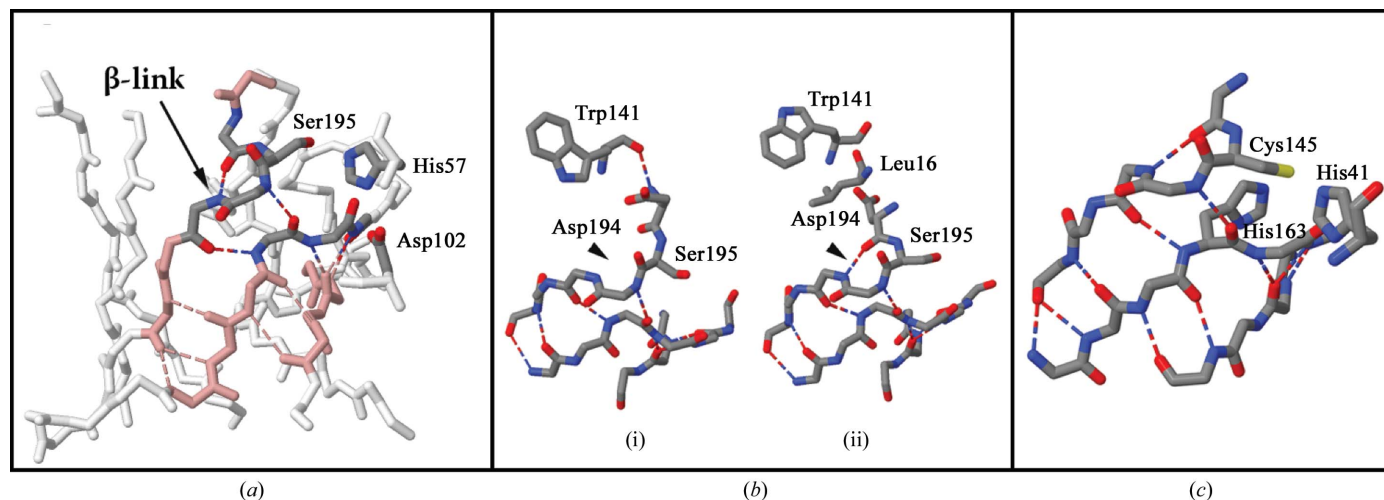


Figure 8

$\beta$ -Links in proteases of SCOPe superfamily b.47.1. (a) Bovine trypsin (PDB entry 1bty; Katz *et al.*, 1995). Backbone plot with the  $\beta$ -link, associated classic  $\beta$ -bulge and active-site residues (labelled and side chains shown) coloured CPK, adjacent residues salmon pink and other parts of the protein white. (b) (i) Human prokallikrein 6 (PDB entry 1gvl chain A; Gomis-Rüth *et al.*, 2002). (ii) Human kallikrein 6 (PDB entry 1l2e chain A; Bennett *et al.*, 2002). Shown is the region near the catalytic Ser195. The arrowheads indicate the position of the hydrogen bond in the  $\beta$ -turn of the  $\beta$ -link found only in the active protein. Side chains are shown for Ser195, Asp194 and residues in proximity to the latter. Leu16 is not shown in (i) because it is distal to the residues shown before the activation of prokallikrein. (c) Human SARS-CoV-2 M<sup>Pro</sup> (PDB entry 7bqy; Jin *et al.*, 2020). The environment of the  $\beta$ -link with the side chains of the active-site Cys145, His163 and His41 is shown.

### 3.5. $\beta$ -Links at the active site of serine and cysteine proteases

No mention has yet been made of a role for the  $\beta$ -link in enzyme function. However, the catalytic Ser195 of serine proteases is located in the  $\beta$ -turn of a  $\beta$ -link, a feature noted by Richardson *et al.* (1978). This is shown for trypsin in Fig. 8(a), where it can be seen that the  $\beta$ -link is at the connection of opposite strands across an end of one of the two six-stranded  $\beta$ -barrels (family b.47.1.2) and that Ser195 occupies the position of residue 0. In all active enzymes of this type examined, including eukaryotic, bacterial and viral serine proteases, the  $\beta$ -link was conserved (Supplementary Table S2). The classic  $\beta$ -bulge (as in Fig. 7b) associated with the  $\beta$ -link is found in the eukaryotic serine proteases examined, but only in some of the bacterial and viral enzymes.

Although the  $\beta$ -link is also present in the inactive precursors of serine proteases, trypsinogen and chymotrypsinogen, it is absent from human prokallikrein 6 (Fig. 8b). The activation of kallikrein differs from that of other serine proteases, as it involves the movement of a new N-terminus (Leu16) to the proximity of Asp194, breaking its interaction with Trp141 and reorienting the nucleophilic Ser195 in position for catalysis (Gomis-Rüth *et al.*, 2002). Fig. 8(b) shows that this also allows the formation of the hydrogen bond of the  $\beta$ -turn of the  $\beta$ -link.

Superfamily b.47.1, containing the serine proteases, also encompasses the superfamily b.47.1.4, which contains viral proteases that are structurally homologous to the serine proteases, but in which the catalytic Ser195 is replaced by a cysteine residue. (Cysteine proteinases such as papain occupy a different fold: d.3.) The catalytic mechanism of these viral proteases also differs from that of the serine proteases in the involvement of a second essential histidine residue (Anand *et al.*, 2003). Both the catalytic cysteine and the second histidine

(His163) are located in the conserved  $\beta$ -link, with the histidine occupying the singleton position (X) of the G1  $\beta$ -bulge component. This is shown in Fig. 8(c) for the protease M<sup>Pro</sup> from human SARS-CoV-2 (Jin *et al.*, 2020). The first histidine, His41, corresponds to the well known His57 of trypsin, but the third component of the catalytic triad in trypsin, Asp102, is absent in the viral protease.

## 4. Discussion

In the original description of the motif studied in this work it was suggested that 'one sort of function' that it might have 'would be to influence the direction in which a strand could leave a  $\beta$ -sheet' (Richardson *et al.*, 1978). The present work describes the architectural nature of such a function. When a strand leaves a  $\beta$ -sheet through a  $\beta$ -link it is frequently directed towards and connects with another structural component of the protein: a second sheet in a sandwich structure, a strand at the other side of a small  $\beta$ -barrel, or an  $\alpha$ -helix. These latter environments are not as different as may at first appear because one can regard many  $\beta$ -barrels as 'two  $\beta$ -sheets packed face to face, with the strands in each sheet lying roughly perpendicular to one another' (Chothia & Janin, 1982; Youkharibache *et al.*, 2019). Furthermore, a short helix is often associated with the  $\beta$ -links in small  $\beta$ -barrels (Youkharibache *et al.*, 2019; Fig. 6). Of course,  $\beta$ -links occur in other contexts, as discussed in Section 3.4, but at least 50% would appear to have this newly identified architectural role.

The specific nature of this is evident from the frequency with which  $\beta$ -links occur at the same position in particular SCOPe folds (Supplementary Table S2), but we recognize that such evolutionary conservation is not absolute. In the  $\beta$ -sandwich we have found that in many cases at positions



where a  $\beta$ -link is not completely conserved a G1 $\beta$   $\beta$ -bulge is still present; for example in family b.1.8.1 (Supplementary Table S2). Here, one supposes that either other features of protein structure fulfil the role previously mentioned for the hydrogen bond of the type II  $\beta$ -turn, or a wider loop between sheets does not require this. In other cases, such as folds b.3 and b.11 (Section 3.2), the connections between the two sheets appear to be facilitated quite differently.

It is possible to imagine that in addition to its role in maintaining a particular protein architecture, the  $\beta$ -link might initiate the folding of the protein into this architecture. If, as is widely, although not universally (Leader & Milner-White, 2011), assumed, proteins fold as they are synthesized, in the N to C direction, this would seem somewhat at odds with position of the  $\beta$ -link on the connection. However, we have no data that bear on this question.

We are not aware of previous reports of the relationship of  $\beta$ -links to architectural classifications such as SCOPe, but Craveur *et al.* (2013) did report such a study of  $\beta$ -bulges. Their work showed high conservation of  $\beta$ -bulges at particular positions, but it did not differentiate between the two subclasses of G1  $\beta$ -bulge, nor consider folds within SCOPe classes, making it difficult to relate to the present study.

Although our emphasis has been on how  $\beta$ -links can connect larger structural components of proteins, we have shown that in at least one circumstance such  $\beta$ -links can play an additional role of their own. The active-site Ser195 of serine proteases, and the cysteine residue of homologous cysteine proteases, are situated at the end of a small  $\beta$ -barrel in which a  $\beta$ -link initiates a loop across the top of the barrel. These residues are located at the same position within the completely conserved  $\beta$ -link, and it seems likely that the extensive inter-main-chain hydrogen-bond network in and around the  $\beta$ -link ensures that the serine or cysteine side chain is orientated optimally for catalysis. The possible relevance of this to the development of inhibitors of viral proteases, such as that of SARS-CoV-2 (Fig. 8c), should not be discounted.

Impressive progress has been made recently in the design and synthesis of proteins, including that of an eight-stranded  $\beta$ -barrel (Dou *et al.*, 2018), but as far as we are aware none of these constructs have yet included  $\beta$ -sandwiches or small  $\beta$ -barrels. We have previously described patterns of amino acids that distinguish  $\beta$ -links from other composites of  $\beta$ -bulges (Leader & Milner-White, 2021) and have also found some small differences between the  $\beta$ -links in  $\beta$ -sandwiches and  $\beta$ -barrels (Supplementary Table S3). Now that the architectural role of the  $\beta$ -link has been recognized, we hope that this will help in the design of increasingly ambitious synthetic proteins.

## References

- Afzal, A. M., Al-Shubailly, F., Leader, D. P. & Milner-White, E. J. (2014). *Proteins*, **82**, 3023–3031.
- Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. (2003). *Science*, **300**, 1763–1767.

- Bernett, M. J., Blaber, S. I., Scarisbrick, I. A., Dhanarajan, P., Thompson, S. M. & Blaber, M. (2002). *J. Biol. Chem.* **277**, 24562–24570.
- Blandl, T., Cochran, A. G. & Skelton, N. J. (2003). *Protein Sci.* **12**, 237–247.
- Bruns, C. M. & Karplus, P. A. (1995). *J. Mol. Biol.* **247**, 125–145.
- Bucher, R. M., Svergun, D. I., Muhle-Goll, C. & Mayans, O. (2010). *J. Mol. Biol.* **401**, 843–853.
- Chan, A. W., Hutchinson, E. G., Harris, D. & Thornton, J. M. (1993). *Protein Sci.* **2**, 1574–1590.
- Chandonia, J.-M., Fox, N. K. & Brenner, S. E. (2017). *J. Mol. Biol.* **429**, 348–355.
- Chattopadhyay, K., Ramagopal, U. A., Mukhopadhyaya, A., Malashkevich, V. N., DiLorenzo, T. P., Brenowitz, M., Nathenson, S. G. & Almo, S. C. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 19452–19457.
- Chothia, C. & Janin, J. (1982). *Biochemistry*, **21**, 3955–3965.
- Craveur, P., Joseph, A. P., Rebehmed, J. & de Brevern, A. G. (2013). *Protein Sci.* **22**, 1366–1378.
- Crowley, P. B., Matias, P. M., Mi, H., Firkbank, S. J., Banfield, M. J. & Dennison, C. (2008). *Biochemistry*, **47**, 6583–6589.
- Davies, G. J., Tolley, S. P., Henrissat, B., Hjort, C. & Schülein, M. (1995). *Biochemistry*, **34**, 16210–16220.
- Dou, J., Vorobieva, A. A., Sheffler, W., Doyle, L. A., Park, H., Bick, M. J., Mao, B., Foight, G. W., Lee, M. Y., Gagnon, L. A., Carter, L., Sankaran, B., Ovchinnikov, S., Marcos, E., Huang, P.-S., Vaughan, J. C., Stoddard, B. L. & Baker, D. (2018). *Nature*, **561**, 485–491.
- Elkins, J. M., Papagrigoriou, E., Berridge, G., Yang, X., Phillips, C., Gileadi, C., Savitsky, P. & Doyle, D. A. (2007). *Protein Sci.* **16**, 683–694.
- Fox, N. K., Brenner, S. E. & Chandonia, J.-M. (2014). *Nucleic Acids Res.* **42**, D304–D309.
- Gomis-Rüth, F. X., Bayés, Á., Sotiropoulou, G., Pampalakis, G., Tsetsenis, T., Villegas, V., Avilés, F. X. & Coll, M. (2002). *J. Biol. Chem.* **277**, 27273–27281.
- Gunasekaran, K., Gomathi, L., Ramakrishnan, C., Chandrasekhar, J. & Balaram, P. (1998). *J. Mol. Biol.* **284**, 1505–1516.
- Herráez, A. (2006). *Biochem. Mol. Biol. Educ.* **34**, 255–261.
- Hutchinson, E. G. & Thornton, J. M. (1994). *Protein Sci.* **3**, 2207–2216.
- Ikemizu, S., Sparks, L. M., van der Merwe, P. A., Harlos, K., Stuart, D. I., Jones, E. Y. & Davis, S. J. (1999). *Proc. Natl Acad. Sci. USA*, **96**, 4289–4294.
- Iyer, L. M. & Aravind, L. (2004). *Proteins*, **55**, 977–991.
- Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., Duan, Y., Yu, J., Wang, L., Yang, K., Liu, F., Jiang, R., Yang, X., You, T., Liu, X., Yang, X., Bai, F., Liu, H., Liu, X., Guddat, L. W., Xu, W., Xiao, G., Qin, C., Shi, Z., Jiang, H., Rao, Z. & Yang, H. (2020). *Nature*, **582**, 289–293.
- Katz, B. A., Finer-Moore, J., Mortezaei, R., Rich, D. H. & Stroud, R. M. (1995). *Biochemistry*, **34**, 8264–8280.
- Ke, H. (1992). *J. Mol. Biol.* **228**, 539–550.
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). *Bioinformatics*, **23**, 2947–2948.
- Leader, D. P. & Milner-White, E. J. (2009). *BMC Bioinformatics*, **10**, 60.
- Leader, D. P. & Milner-White, E. J. (2011). *Proteins*, **79**, 1010–1019.
- Leader, D. P. & Milner-White, E. J. (2012). *BMC Struct. Biol.* **12**, 26.
- Leader, D. P. & Milner-White, E. J. (2015). *Proteins*, **83**, 2067–2076.
- Leader, D. P. & Milner-White, E. J. (2021). *Acta Cryst. D* **77**, 217–223.
- Leonidas, D. D., Elbert, B. L., Zhou, Z., Leffler, H., Ackerman, S. J. & Acharya, K. R. (1995). *Structure*, **3**, 1379–1393.
- Machius, M., Cianga, P., Deisenhofer, J. & Ward, E. S. (2001). *J. Mol. Biol.* **310**, 689–698.
- Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998). *Nat. Struct. Mol. Biol.* **5**, 721–729.
- McDonald, I. K. & Thornton, J. M. (1994). *J. Mol. Biol.* **238**, 777–793.
- Milner-White, E. J. (1987). *Biochim. Biophys. Acta*, **911**, 261–265.

- Milner-White, E. J. (1988). *J. Mol. Biol.* **199**, 503–511.
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.
- Peat, T. S., Newman, J., Waldo, G. S., Berendzen, J. & Terwilliger, T. C. (1998). *Structure*, **6**, 1207–1214.
- Rabijns, A., De Bondt, H. L. & De Ranter, C. (1997). *Nat. Struct. Mol. Biol.* **4**, 357–360.
- Richardson, J. S. (1981). *Adv. Protein Chem.* **34**, 167–339.
- Richardson, J. S., Getzoff, E. D. & Richardson, D. C. (1978). *Proc. Natl Acad. Sci. USA*, **75**, 2574–2578.
- Schellmann, J. A. (1980). *Protein Folding*, edited by R. Jaenicke, pp. 53–61. Amsterdam: Elsevier.
- Venkatachalam, C. M. (1968). *Biopolymers*, **6**, 1425–1436.
- Watson, J. D. & Milner-White, E. J. (2002). *J. Mol. Biol.* **315**, 171–182.
- Wiegand, G., Epp, O. & Huber, R. (1995). *J. Mol. Biol.* **247**, 99–110.
- Wilmot, C. M. & Thornton, J. M. (1988). *J. Mol. Biol.* **203**, 221–232.
- Yokoyama, S., Kigawa, T., Shirouzu, M., Miyano, M. & Kuramitsu, S. (2008). *Tanpakushitsu Kakusan Koso*, **53**, 632–637.
- Youkharibache, P., Veretnik, S., Li, Q., Stanek, K. A., Mura, C. & Bourne, P. E. (2019). *Structure*, **27**, 6–26.